MOX-Report No. 99/2024

Analysis of Higher Education Dropouts Dynamics through Multilevel Functional Decomposition of Recurrent Events in Counting Processes

Ragni, A.; Masci, C.; Paganoni, A. M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it https://mox.polimi.it

Analysis of Higher Education Dropouts Dynamics through Multilevel Functional Decomposition of Recurrent Events in Counting Processes

Alessandra Ragni*, Chiara Masci*, Anna Maria Paganoni*
*MOX, Department of Mathematics, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133, Milano, Italy

Abstract

This paper analyzes the dynamics of higher education dropouts through an innovative approach that integrates recurrent events modeling and point process theory with functional data analysis. We propose a novel methodology that extends existing frameworks to accommodate hierarchical data structures, demonstrating its potential through a simulation study. Using administrative data from student careers at Politecnico di Milano, we explore dropout patterns during the first year across different bachelor's degree programs and schools. Specifically, we employ Cox-based recurrent event models, treating dropouts as repeated occurrences within both programs and schools. Additionally, we apply functional modeling of recurrent events and multilevel principal component analysis to disentangle latent effects associated with degree programs and schools, identifying critical periods of dropout risk and providing valuable insights for institutions seeking to implement strategies aimed at reducing dropout rates.

Keywords: students dropout, recurrent events, multilevel principal component analysis, functional data analysis

1 Introduction

The higher education system is worldwide affected by high dropout rates. In this context, "dropout" refers to students leaving the university world without completing their degree. From the perspective of a single university, dropout occurs when a student exits their academic program before earning the final qualification (Tinto, 1982). Despite efforts by European governments to expand access to higher education, ensuring successful degree completion remains a challenge, and dropout rates persist at around 30% across OECD member countries (OECD, 2019). In Italy, this issue is particularly pronounced, with a significant proportion of students discontinuing their studies, often within the first two years of enrollment. More than half of those who begin higher education fail to complete their degrees (Aina et al., 2018). Indeed, the percentage of adults with a higher education degree in Italy is below the OECD average (OECD, 2019; Cannistrà, 2024). These high dropout rates not only lower the average skill levels of the workforce (Atzeni et al., 2022), but they are also linked to a growing wage-skill gap (Katz and Murphy, 1992).

From an institutional perspective, high dropout rates represent a waste of resources. In fact, the long-term returns — both in terms of human capital development and the credentials awarded — are lost when students exit without completing their degrees, despite the considerable investments made by universities in teaching, recruitment, and student support. As a result, reducing and analyzing university dropout rates has become a critical challenge for higher education institutions.

What makes managing this issue even more complex is the significant variation in dropout behavior across degree programs and schools. Even within the same university, dropout patterns differ widely across academic disciplines. For instance, some programs may experience higher dropout rates during early semesters due to difficult coursework, while others might see students leave later in their studies, near graduation. Additionally, dropout rates can vary between schools within the same univesity, influenced by factors such as faculty engagement, availability of student support services, and workload.

In this paper, we analyze administrative data from Politecnico di Milano (PoliMi) to examine dropout patterns across its bachelor's degree programs. PoliMi comprises four distinct schools: Architecture,

Design and Engineering, further divided into the School of Civil, Environmental, and Land Management Engineering and the School of Industrial and Information Engineering, offering 23 different undergraduate programs, referred to as degree courses or simply courses. Our focus is on understanding dropout rates across these programs, exploring how they vary by both degree program and school over the first-year span.

Our approach builds on methodologies that integrate recurrent events modelling, point process theory, and functional data analysis, extending the techniques proposed by Baraldo et al. (2013) and Spreafico and Ieva (2021). In these studies, hospital readmissions and drug consumption over time are analyzed to predict outcomes related to heart failure telemonitoring in the former and time-to-death in the latter. We extend and generalize this framework to account for the hierarchical structure of the data (Pinheiro and Bates, 2000), enabling a more detailed exploration of dropout dynamics across different academic units. Specifically, the analysis comprises two phases. In the first phase, we utilize historical dropout data to fit a counting process model (Daley and Vere-Jones, 2002), enabling us to compute the realized trajectories of the cumulative hazard process (*compensators)* underlying the dropout counting process. While many alternatives are available for the modelization of recurrent events (Amorim and Cai, 2015), such as extensions of Cox models (see, for instance, the Prentice, Williams and Peterson model (Prentice et al., 1981), Wei, Lin and Weissfeld model (Wei et al., 1989), frailty models (Therneau et al., 2000)), models for the mean number of events or their occurrence rate (Lin et al., 2000; Diao et al., 2014), multi-state models (Andersen and Keiding, 2002), and virtual (effective) age models (Kijima et al., 1988; Peña and Hollander, 2004; Beutner, 2023), we employ the Andersen-Gill (AG) model (Andersen and Gill, 1982). This choice follows Spreafico and Ieva (2021), the most recent research in this context. The AG model extends the Cox proportional hazards model by incorporating the increments in event counts over time, assuming that correlations between event times can be explained by prior occurrences, as well as through the specification of appropriate time-varying covariates, such as the count of previous occurrences (Amorim and Cai, 2015). This allows us to represent these events as non-stationary stochastic counting processes that may depend on specific characteristics or labels, referred to as *marks* (Daley and Vere-Jones, 2002; Spreafico and Ieva, 2021). At this stage, the longitudinal trajectory of instantaneous dropout risk over time within a degree program is treated as a function, and functional data analysis techniques (Ramsay and Silverman, 2005) are employed to extract insights from repeated dropout events as two-level functional covariates. These covariates are derived through dimensionality reduction using Multilevel Functional Principal Component Analysis (MFPCA) (Di et al., 2009; Cui et al., 2023), preserving most of the historical information while effectively managing variability across two levels. In this first phase, our aims are twofold: (i) to reconstruct the dropout curve by modeling dropout intensity as a counting process, capturing the temporal dynamics of dropout risk in terms of cumulative hazard for each degree program, and (ii) to decompose this dropout curve into contributions from both the degree program and school levels using MFPCA, highlighting how different academic units influence the evolution of dropout risk. In the second phase, we adopt a predictive framework to investigate how these covariates influence the subsequent risk of dropout among students, incorporating information specific to the dropout risk associated with each faculty or school. The aim of this second phase is to assess the predictive value of the extracted functional covariates in forecasting future dropout events at the student level.

Previous studies focused on PoliMi data have addressed various aspects of student dropout prediction and quantification (Cannistrà et al., 2022; Romani, 2023; Diaz Lema et al., 2024; Masci et al., 2024), with some of them specifically examining the impact of grouping factors — such as degree programs — on the time to dropout within the first few semesters of enrollment up to the full three years of bachelor degree. The tools commonly employed in this context include shared frailty Cox proportional hazard models (Cook et al., 2007; Kleinbaum and Klein, 1996), where the frailty term represents a constant factor shared among clusters (e.g., degree programs), which affects the baseline hazard multiplicatively, accounting for unobserved heterogeneity within clusters and allowing for a more detailed understanding of the dropout risk across different academic programs. Our study extends previous research by offering a more refined analysis of dropout behaviour over time, specifically examining how dropout dynamics evolve both within and between degree programs and schools. A key advancement is the incorporation of the dropout history into the predictive framework which, unlike the shared frailty model that simplifies this information into a single measure, allows for a more detailed and interpretable analysis of dropout patterns, offering institutions tools for developing targeted strategies aimed at reducing dropout rates. In this perspective, two key elements of novelty need to be highlighted. First, we introduce the use of FMPCA to decompose the dropout curve constructed from the compensator function. This novel application allows us to separate the contributions of different academic units (e.g., programs and schools) to the overall dropout dynamics, providing a richer understanding of how these hierarchical factors influence dropout risk over time. Second, and more importantly, our approach represents a completely

new perspective in the dropout literature, where most studies focus on classification-based predictive models or, in more sophisticated cases, time-to-event models, almost exclusively at the student level (see, for instance, Arulampalam et al. (2004); Plank et al. (2008); Min et al. (2011); Gury (2011); Vallejos and Steel (2017); Patacsil (2020), and Masci et al. (2024) for a discussion). In contrast, our approach models dropout dynamics at the degree course level, capturing historical temporal patterns that can later be included for predictions at student level.

The paper is structured as follows. In Section 2 we introduce the PoliMi dataset, detailing the cohort selection and study design. Section 3 outlines the employed methodology, providing a recap of the framework and extending it within a multilevel context. Section 4 reports on a simulation study that illustrates the application of the proposed methodology within the multilevel framework. Section 5 presents the results obtained from applying the proposed methodology to the PoliMi case study. Discussion and concluding remarks are provided in Section 6.

2 Dataset

The data employed in this study were obtained from the administrative records of PoliMi, which collect the academic progress of students enrolled in bachelor's degree programs (Mussida and Lanzi, 2022). These records encompass various aspects of students' academic *careers*, including enrollment and end-of-study dates, any changes in their enrolled degree programs, and eventually incidents of dropout. Additionally, information regarding the student's history of passed and attempted *exams* at different time points (semesters) is contained, including credits earned within the European Credit Transfer and Accumulation System (ECTS) and weighted Grade Point Average (GPA). In this section, we delineate the cohort selection criteria for our study (Subsections 2.1) alongside the study design (Subsection 2.2).

2.1 Cohort selection

For our analysis, we focus on bachelor's students enrolled in academic years 2016/2017 and 2017/2018, who maintained a consistent degree program throughout their academic paths. We assume that students enrolled in the 2017/2018 academic year were only marginally impacted by Covid-19, which occurred during the last semester of their final year.

We exclude students who graduated in under 1000 days (the minimum duration for a bachelor's degree at PoliMi) and omit fully remote or single-cycle degree programs, as the analysis focuses solely on traditional bachelor's degrees.

In the first phase of the analysis, we utilize data from students with `career_start_ay` = '2016' to construct the compensators. For these students, we track the dropout events occurring within each `course` and `school` during the first three semesters since enrollment. In the second phase, we shift to the 2017 cohort, using data from the end of the first semester and historical information to enhance dropout risk predictions. Table 1 provides an overview of the key variables used in the second phase, organized into four categories:

- Variables measured at enrollment capture essential demographic and background characteristics. These include geographic origin (`origins`, i.e., whether a student lives onsite, offsite, or commutes to Milan), gender (`gender`), and age at enrollment (`age19`, which identifies students older than 19). Socio-economic status is approximated by the university fee bracket (`income`), which classifies students based on their family's financial situation into categories such as low, medium, high, or those receiving grants. Educational background is represented by the type of high school attended (`highschool_type`). Lastly, the PoliMi admission test score (`admission_score`) reflects academic readiness at the time of university entry, although students may take this test up to a year prior to their actual enrollment.
- Variables measured at the end of the first semester focus on academic progress, particularly the number of credits earned (`ECTS1sem`), a key predictor of dropout risk.
- Grouping factors include the undergraduate program (`course`) and broader organizational structure (`school`).
- The outcome variable (`dropout3y`) indicates whether a student dropped out within three years, after the first semester.

Variable	Description	Type
Measured at enrollment		
- `studentID`	Student's unique identifier (anonymized)	Categorical {`1,2,...`}
- `origins`	Student's geographic origins	Categorical {`OnSite, Commuter, Offsite`}
- `gender`	Student's gender	Categorical {`Male, Female`}
- `highschool_type`	Type of attended high school	Categorical {`Scientific, Classical, Others, Technical`}
- `income`	University fee brakcet	Categorical {`Medium, Grant, High, Low`}
- `age19`	Equals 1 if student's age at enrollment > 19	Categorical {`0,1`}
- `admission_score`	PoliMi entrance test's admission score	Real number [60, 100]
- `career_start_ay`	Student's enrollment year	Categorical {`2016, 2017`}
Measured at end of 1st semester		
- `ECTS1sem`	ECTS gained by end of 1st semester	Natural number > 0
Grouping factors		
- `course`	Undergraduate program	Categorical {`P01, P02,..., P23`}
- `school`	A larger organizational unit grouping courses	Categorical {`sA, sB, sC, sD`}
Outcome		
- `dropout3y`	Equals 1 if after 1st semester a student drops within 3 years, 0 otherwise	Categorical {`0, 1`}

Note: in categorical variables, the first reported class represents the reference level.

Table 1: Overview of the variables considered in the analysis.

2.2 Study design

We label as *dropouts* the students who dropped out between the end of the first and sixth semester (*follow-up*), while as *censored* all the other students that dropped out after three years from the enrollment, who graduated or who had an active career at the end of the third year.

With focus on students enrolled in `career_start_ay` = '2016', we include a three-semesters *observation period*, denoted by $S = [T_0, T_1]$ with T_0 = '`career_start_ay`/10/01' and T_1 = '`career_start_ay` + 2/03/01'. The date of October 1st is chosen to exclude students who dropped out within the first two weeks, potentially due to waiting for other university entrance test results[1], to ensure they do not affect the analysis. During this period, dropouts of students enrolled in the chosen `career_start_ay` are monitored and analyzed based on various grouping factors (`course` and `school`).

Following this, the focus is moved to the cohort of students with `career_start_ay` = '`2017`', particularly at the end of the first semester, and the primary outcome of interest is the binary variable `dropout3y` indicating whether a student dropped out within three years from enrollment. To predict this outcome, we use data collected at enrollment, at the end of the first semester, and at the level of grouping factors. The choice is based on previous studies' results (Masci et al., 2024), which indicate that the optimal prediction window occurs within the first few semesters, as the inclusion of data from later semesters provides minimal improvement in accuracy. Moreover, incorporating hierarchical information has been shown to enhance predictive performance, with the number of credits earned by the end of the first semester serving as a particularly strong predictor. Furthermore, the model incorporates information derived from the analysis of the previous academic year's data, adding valuable historical context to enhance predictive accuracy.

3 Methodology

In this section, we present the methodology in three consequent steps: recap on model formulation for recurrent events and compensators reconstruction (Subsection 3.1), compensators decomposition through multilevel principal component analysis (Section 3.2) and the development of a predictive model for the dropout status within three years including retrieved information (Section 3.3). The core methodological contribution of this work regards the extension to the multilevel setting of the decomposition of recurrent events in counting processes.

[1] At PoliMi, lectures typically begin in mid-September.

3.1 Recap on the model formulation for recurrent events and compensators reconstruction

Let $N_{ij}(t)$, with $t \in [0, T]$, denote the stochastic process counting the dropout events observed up to time t, where $j = 1, ..., J_i$ indexes the `course`-level (lower-level) units and $i = 1, ..., I$ indexes the `school`-level (higher-level) units, or clusters, with the total number of lower-level units given by $\sum_{i=1}^{I} J_i = n$ (Cook et al., 2007). The process $N_{ij}(t)$ is adapted to the filtration $\{\mathcal{F}_{t,ij}\}_{t\in[0,T]}$, that is the history of realizations of the process itself. Assuming $N_{ij}(t)$ is a class D submartingale, the Doob-Meyer (D-M) decomposition theorem (Meyer, 1962) states that $M_{ij}(t) = N_{ij}(t) - \Lambda_{ij}(t)$ is a zero-mean, uniformly integrable martingale. Here, $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(s)ds$ is the unique predictable non-decreasing cadlag[2] and integrable *compensator* (or cumulative hazard process), with $\lambda_{ij}(t)$ being the intensity process (or hazard function).

Building on the formulation for marked point processes described in Spreafico and Ieva (2021) and extending it to a multilevel context, the events, whose cumulative number up to a given time t are recorded by the counting process $N_{ij}(t)$, can be further associated to additional random variables (marks) $\boldsymbol{\omega}_{ij}$ that provide further details about these events, such as the size or magnitude related to the jumps in the counting process (Last and Brandt, 1995; Daley et al., 2003). In this framework, the conditional intensity function also depends on the mark $\boldsymbol{\omega}_{ij}$. Assuming conditional independence of jump times and marks, the following relationship holds $\lambda_{ij}(t, \boldsymbol{\omega}_{ij}) = \lambda_{ij,g}(t) \cdot f_{ij}(\boldsymbol{\omega}_{ij})$, where $\lambda_{ij,g}(t)$ is the ground intensity process of the counting process and f_{ij} is the multivariate density of the marks $\boldsymbol{\omega}_{ij}$. Proper modeling of compensators and particularly of $\lambda_{ij}(t, \boldsymbol{\omega}_{ij})$, allows for an accurate reconstruction of $N_{ij}(t)$, as $M_{ij}(t)$ represents the residual of the process in the D-M decomposition.

Several models for $\lambda_{ij}(t)$ are available in the literature on counting processes (Aalen et al., 2008; Andersen et al., 2012; Peña and Hollander, 2004). Employing the model introduced by (Andersen and Gill, 1982), under the assumption that $f_{ij}(\boldsymbol{\omega}_{ij})$ depends on $\mathbf{z}_{ij}(t)$ (that are some time-dependent features related to the marks $\boldsymbol{\omega}_{ij}$), we get

$$\begin{aligned}\lambda_{ij}(t, \boldsymbol{\omega}_{ij}) &= Y_{ij}(t)\ \lambda_0(t)\ \exp\{\boldsymbol{\beta}^T \mathbf{x}_{ij}(t)\}\ \exp\{\boldsymbol{\theta}^T \mathbf{z}_{ij}(t)\} \\ &= Y_{ij}(t)\ \lambda_0(t)\ \exp\{\boldsymbol{\beta}^T \mathbf{x}_{ij}(t) + \boldsymbol{\theta}^T \mathbf{z}_{ij}(t)\} \end{aligned} \tag{1}$$

where $\mathbf{x}_{ij}(t)$ are the (time-dependent) column vectors of covariates of the j^{th} unit in i^{th} cluster, $\lambda_0(t)$ is the baseline hazard function, Y_{ij} takes the role of the censoring variable (i.e. assumes value 1 when unit j in cluster i is under observation), $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are Q- and P-dimensional column vectors of coefficient and ^{T} stands for the transpose. In particular, following Spreafico and Ieva (2021), the mark density $f_{ij}(\boldsymbol{\omega}_{ij})$ is incorporated into the model through the exponential term involving $\mathbf{z}_{ij}(t)$, which parametrizes the influence of the marks on the process. $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated in the model fitting by partial likelihood maximization (Andersen and Gill, 1982), while the baseline cumulative hazard $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ can be estimated through Breslow estimator (Breslow, 1975) as a step-function $\hat{\Lambda}_0(t)$ and then smoothed into $\tilde{\Lambda}_0(t)$ as described in Baraldo et al. (2013).

Let now $[t_k^{(ij)}, t_{k+1}^{(ij)}]$ for $k = 0, ..., N_{ij}(T)$ be the intervals whose extremes are the jump times for each unit j in cluster i, being $t_0^{(ij)} = 0$ and $t_{N_{ij}(T)+1}^{(ij)} = T$. Then $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(s, \boldsymbol{\omega}_{ij})ds$ can be estimated by approximation as follows (see computation in Appendix A):

$$\hat{\Lambda}_{ij}(t) = \sum_{k=0}^{N_{ij}(t^-)} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}(t_k^{(ij)}) + \hat{\boldsymbol{\theta}}^T \mathbf{z}_{ij}(t_k^{(ij)}))\big[\tilde{\Lambda}_0(t_{k+1}^{(ij)} \wedge t) - \tilde{\Lambda}_0(t_k^{(ij)})\big]. \tag{2}$$

where $a \wedge b = \min\{a, b\}$, $N_{ij}(t^-)$ represents the number of occurrences that have happened strictly before time t, and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are the estimated vectors of coefficients.

3.2 Multilevel functional principal component analysis for compensators decomposition

After reconstructing compensators through a marked point process formulation for recurrent events, $\hat{\Lambda}_{ij}(t)$ can be regarded as functional data objects, allowing the application of functional data analysis techniques (Ramsay and Silverman, 2005).

Given the high-dimensional nature of these data and the hierarchical setting, we aim to decompose functional variability and reduce dimensionality, while getting insights. To achieve this, we apply MFPCA

[2] i.e., right-continuous with left limits.

(Di et al., 2009; Cui et al., 2023). MFPCA integrates classical FPCA (Ramsay and Silverman, 2005), which selects only the relevant components of an appropriate orthonormal basis expansion, with standard multilevel mixed models. This approach effectively decomposes the observed data according to two levels of functional variation. Specifically, from the one-way functional ANOVA (Di et al., 2009) follows that

$$\hat{\Lambda}_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t) + \epsilon_{ij}(t) \tag{3}$$

$$= \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k^{(1)}(t) + \sum_{l=1}^{\infty} \zeta_{ijl}\phi_l^{(2)}(t) + \epsilon_{ij}(t) \tag{4}$$

where, in Eq. (3), $\mu(t)$ is a fixed functional effect, $Z_i(t)$ and $W_{ij}(t)$ are mean 0 stochastic processes (uncorrelated between each other) and ϵ_{ij} is observed only when functional data are observed with errors. Eq. (4) follows from Karhunen-Loève (KL) expansion (Karhunen, 1947; Loeve, 1948), where $\phi_k^{(1)}(t)$ and $\phi_l^{(2)}(t)$ are respectively level 1 (i.e., cluster level) and level 2 (i.e., unit level) eigenfunctions (fixed functional effects), and ξ_{ik} and ζ_{ijl} are respectively level 1 and 2 principal component scores (zero mean random variables, uncorrelated between each other). Moreover, one may truncate the decomposition by pre-specifying at both levels the Percentage of Variance Explained (PVE) as explained in Di et al. (2009), resulting into

$$\hat{\Lambda}_{ij}(t) \simeq \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k^{(1)}(t) + \sum_{l=1}^{L} \zeta_{ijl}\phi_l^{(2)}(t) + \epsilon_{ij}(t) \tag{5}$$

being K and L the number of principal components finally identified respectively at level 1 (cluster) and level 2 (unit). Other interesting indicators defined in Di et al. (2009) are the total explained variance between-clusters and within-clusters, and the proportion of variability explained by level 1.

3.3 Logistic regression model with functional compensators

The compensators decomposition in previous sections allows to extract dropout information focusing on the *observation period* $S = [0, T]$, where the units are at `course`-level j, clustered within `school`-level i. As a last step, we include this functional information into a logistic regression model which considers another cohort, where the units are now at `studentID`-level h, for $h = 1, ..., H_{ij}$, nested within level j (`course`-level) again nested within level i (`school`-level), so that $n_{\text{tot}} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} H_{ij}$ is the total number of units. Let $Y_{ijh} \sim \text{Bernoulli}(p_{ijh})$ be the binary variable indicating the output `dropout3y`. Then

$$\begin{aligned}
\text{logit}(p_{ijh}) &= \boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \int_S \Lambda_{ij}(s)\alpha(s)ds \\
&\simeq \boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \int_S \Big[\sum_{k=1}^{K} \xi_{ik}\phi_k^{(1)}(s) + \sum_{l=1}^{L} \zeta_{ijl}\phi_l^{(2)}(s)\Big]\alpha(s)ds \\
&= \boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \sum_{k=1}^{K} \xi_{ik}\alpha_k^{(1)} + \sum_{l=1}^{L} \zeta_{ijl}\alpha_l^{(2)}
\end{aligned} \tag{6}$$

for $i = 1, ..., I$, $j = 1, ..., J_i$ and $h = 1, ...H_{ij}$ where $\boldsymbol{\gamma}$ is a q-dimensional vector of parameters to be estimated, $\mathbf{w}_{ijh}$ is a vector of covariates available at unit level h, $\alpha : S \to \mathbb{R}$ is a functional parameter and $\text{logit}(x) := \ln\left(\frac{x}{1-x}\right)$. The second line follows from Eq. (5) and last equality is given by rewriting $\alpha(s)$ according to different representations into the two different orthonormal bases $\phi_k^{(1)}$ and $\phi_l^{(2)}$, thanks to the orthonormality property; the subscripts are added in order to distinguish the two projections.

4 Simulation Study

In this section, we aim to demonstrate the effectiveness of the methodology described above, in particular in Sections 3.1 and 3.2. After simulating unit-level intensities with shapes based on specific similarities within clusters and generating the Non-Homogeneous Poisson Processes (NHPPs) from these intensities, we show that compensators reconstruction using AG models effectively recovers the within-cluster similarities.

Specifically, in Subsection 4.1, we begin by simulating intensities $\lambda_{ij}(t)$ following a similar methodology to Cui et al. (2023); Di et al. (2009). We employ a one-way functional ANOVA model to capture

similarities within clusters and integrate $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du$ to obtain compensator-like shapes, translating the intensity functions into cumulative hazard functions, and we simulate event times from the $\lambda_{ij}(t)$. In Subsection 4.2.1, we fit AG models to the simulated event data, and reconstruct the compensators $\hat{\Lambda}_{ij}(t)$. Finally, in Subsection 4.2.2 we evaluate the consistency of the information captured by MFPCA before and after NHPPs extraction. We aim to show that cumulative hazard reconstruction using AG models preserves the essential information captured by the MFPCA.

4.1 Data Generating Process

Let $\lambda_{ij}(t)$ be an intensity function measured over a continuous variable $t \in [0, 1]$ for observation j within cluster i, for $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$, generated by a modified one-way functional ANOVA model (Morris et al., 2003) as follows:

$$\lambda_{ij}(t) := \mu(t) + 2 \cdot i \cdot \Big(\sum_{k=1}^{K} \xi_{ik}\, \phi_k^{(1)}(t) + \sum_{l=1}^{L} \zeta_{ijl}\, \phi_l^{(2)}(t) + \epsilon_{ij}(t) \Big) \tag{7}$$

where $\mu(t) = 200$, $\xi_{ik} \sim \mathcal{N}(0, \lambda_k^{(1)})$, $\zeta_{ijl} \sim \mathcal{N}(0, \lambda_l^{(2)})$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

For our data generation, we draw inspiration from the simulation study described in Section 4 of Di et al. (2009), introducing a few modifications to make the generation of the intensities more suitable for our specific context. Firstly, we include a constant $\mu(t)$ to increase the frequency of events in the NHPP generated by $\lambda_{ij}(t)$. Additionally, we scale level 1 and 2 components by a cluster-dependent constant, enhancing the differentiation between-groups and, consequently, the cumulative intensities. Lastly, we assign a higher standard deviation to the true eigenvalues at level 1 compared to level 2, further distinguishing clusters and reducing within-cluster variability.

The decision to simulate the intensities rather than directly simulating the cumulative hazard function stems from the specific characteristics required for the cumulative hazard (increasing monotonicity and ensuring that $\Lambda_{ij}(0) = 0$). Simulating scores from a normal distribution while maintaining these properties is not possible. Therefore, we opt to simulate the intensities to ensure these essential characteristics are preserved.

We assume $I = 20$ clusters, $J = 4$ units and $K = L = 4$. The chosen value of the eigenvalues are $\lambda_k^{(1)} = 0.9^{k-1}$ for $k = 1, ..., K$ and $\lambda_l^{(2)} = 0.2^{l-1}$ for $l = 1, ..., L$, while the chosen value of the eigenfunctions, chosen following Di et al. (2009); Cui et al. (2023), are

$$\{\phi_1^{(1)}(t), \phi_2^{(1)}(t), \phi_3^{(1)}(t), \phi_4^{(1)}(t)\} = \{\sqrt{2}\sin(2\pi t), \sqrt{2}\cos(2\pi t), \sqrt{2}\sin(4\pi t), \sqrt{2}\cos(4\pi t)\} \tag{8}$$

$$\{\phi_1^{(2)}(t), \phi_2^{(2)}(t), \phi_3^{(2)}(t), \phi_4^{(2)}(t)\} = \{1, \sqrt{3}\,(2t-1), \sqrt{5}\,(6t^2-6t+1), \sqrt{7}\,(20t^3-30t^2+12t-1)\} \tag{9}$$

at levels 1 and 2, respectively. Moreover, we assume $\mu(t) = 100$ and $\sigma = 0$ (no noise). Afterwards, we compute the cumulative hazard function as $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du$. In Figure 1, we illustrate the simulated intensities and cumulative hazards.

Following Pasupathy (2010), through the thinning method we simulate event times for a NHPP over $[0, 1]$, where the process is characterized by the time-varying intensity function $\lambda_{ij}(t)$.

4.2 Results

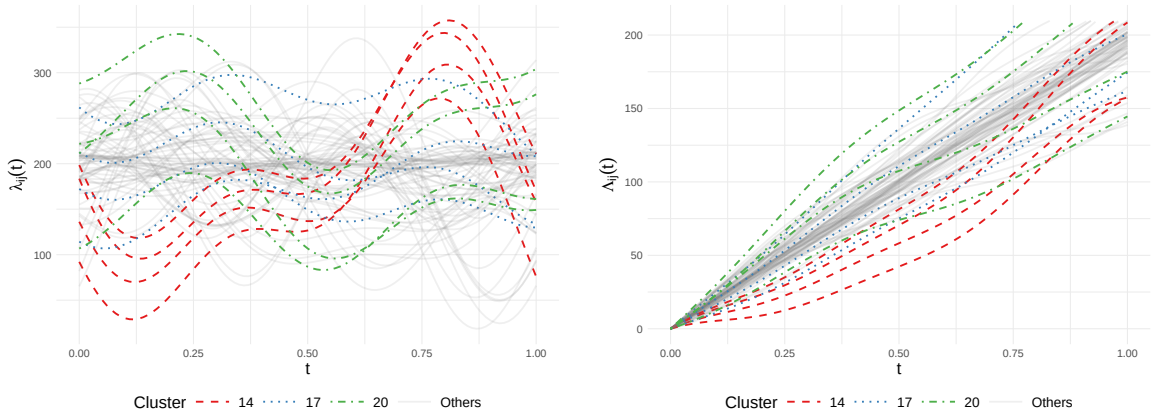
4.2.1 Compensators estimation

At this point, we apply the pipeline described in Section 3.1, where in the AG model in Eq. (1) we employ the number of events recorded up to that time interval for unit j in cluster i as a time-dependent covariate.

The baseline cumulative hazard, $\hat{\Lambda}_0(t)$, along with $\tilde{\Lambda}_0(t)$ is estimated. Additionally, $\hat{\Lambda}_{ij}(t)$ is derived according to Eq. (2). These functions are depicted in Figure 2. By visual inspection, we observe some information loss due to the stochastic nature of NHPPs in the simulation of event times. Nonetheless, cluster behaviours remain distinguishable and can still be effectively recognised and characterized.

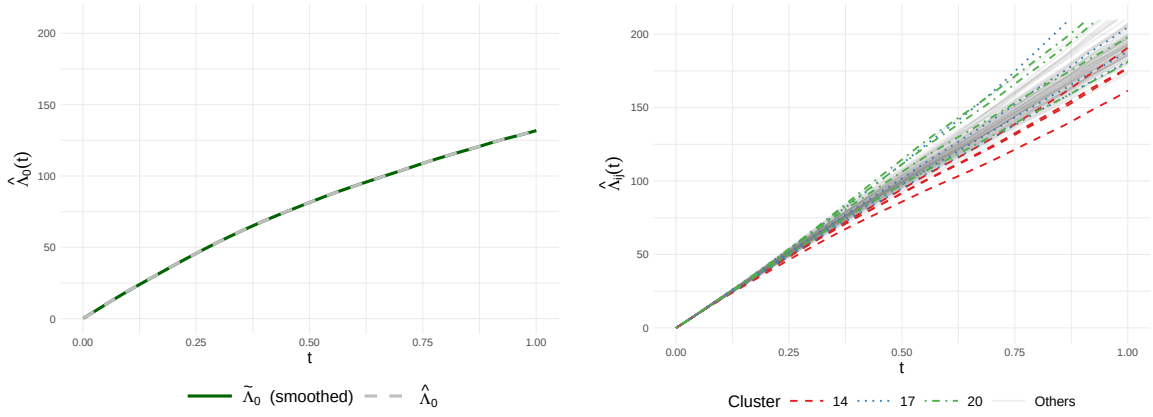
4.2.2 Multilevel functional principal component analysis

As final step of our simulation study, we implement the decomposition described in Section 3.2. We recall that in Eq. (7) we simulate intensities $\lambda_{ij}(t)$ employing the eigenfunctions in Eq. (8-9). However, our primary interest lies in $\Lambda_{ij}(t)$. Therefore, we first decompose $\Lambda_{ij}(t)$ using Eq. (5). At this stage,



(i) $\lambda_{ij}(t)$ (ii) $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du$

Figure 1: Simulated intensities $\lambda_{ij}(t)$ (i) and cumulative hazards $\Lambda_{ij}(t)$ (ii), with clusters 14, 17, and 18 highlighted using different colors and line types due to their outlying shapes, which stand out from the general patterns observed in other clusters.



(i) $\hat{\Lambda}_0(t)$ (ii) $\hat{\Lambda}_{ij}(t)$

Figure 2: Estimated $\hat{\Lambda}_0(t)$ (i) and $\hat{\Lambda}_{ij}(t)$ (ii). In (ii), clusters 14, 17, and 18 are highlighted using different colors and line types due to their outlying shapes, which stand out from the general patterns observed in other clusters.

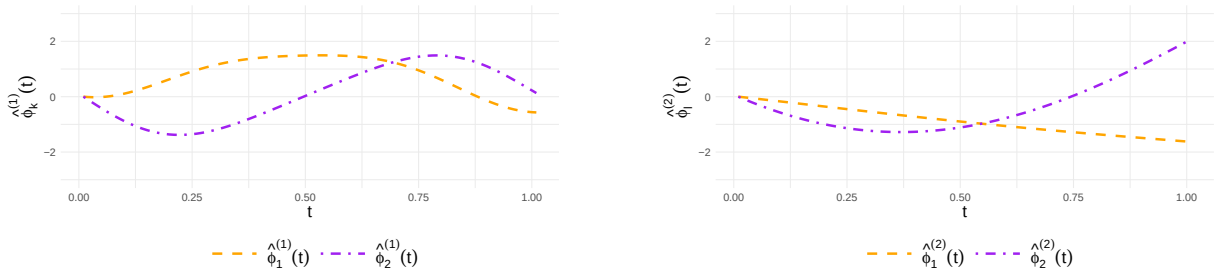


Figure 3: First two eigenfunctions at levels 1 and 2 (left and right panels, respectively), computed from simulated $\Lambda_{ij}(t)$ represented in Figure 1 (ii).

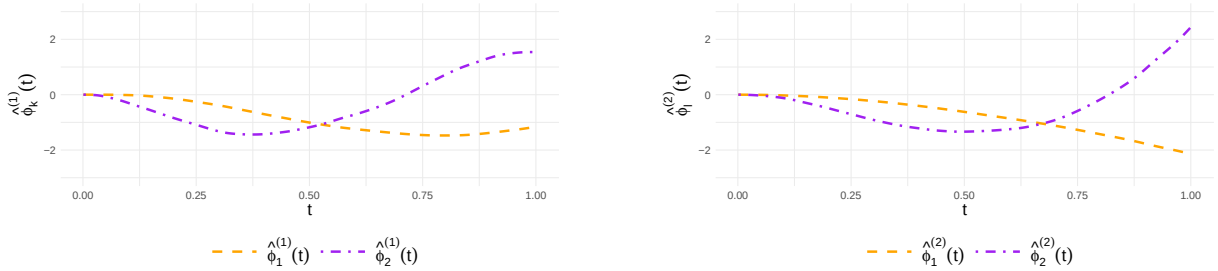


Figure 4: First two eigenfunctions at levels 1 and 2 (left and right panels, respectively), computed from reconstructed $\hat{\Lambda}_{ij}(t)$ represented in Figure 2 (ii).

we obtain 4 functional principal components at level 1 and 2 at level 2. To determine the number of principal components at both levels, we set the PVE to 0.99, following the default setting in the `mfpca.face` function from Cui et al. (2023). In Figure 3, we show results of MFPCA on functional compensators related to the first and second principal components, respectively for levels 1 and 2.

On the other hand, after the simulation of the NHPP as described in previous section and having computed $\hat{\Lambda}_{ij}(t)$ according to Eq. (2), we apply the same multilevel functional decomposition to $\hat{\Lambda}_{ij}(t)$. Here, we obtain 3 functional principal components at level 1 and 2 at level 2 and of degree course, the magnitude of the eigenvalues reduce. However, if we analyse the eigenfunctions reported in Figure 4, similar pattern can be observed, both for levels 1 and 2.

In general, we observe that the reconstructed compensators closely match the original simulated shapes, despite some information loss due to the sampling and fitting processes. Results indicate that cumulative hazard reconstruction using AG models preserves the essential information captured by MFPCA and could be further enriched by incorporating additional application-specific covariates. This demonstrates that for a NHPP with cluster-similar intensities, the cumulative hazard reconstruction using AG models effectively retrieves the simulated shapes, retaining the crucial information captured by MFPCA prior to process extraction.

5 Case Study

We apply the proposed methodology to a case study involving the administrative dataset of PoliMi. First, we present the compensators reconstruction and decomposition at degree `course` and `school` levels (Subsection 5.1), then effectively implement a predictive model at `studentID`-level (Subsection 5.2).

5.1 Compensators reconstruction and decomposition

For the analysis of dropouts as a marked point process, after cohort selection described in Subsection 2.1 and having filtered the data to focus on a specified academic year (in our case, `career_start_ay` = '2016'), we establish start and stop dates for each dropout event that happened on distinct days and enumerate the cumulative occurred dropout distinct days (`enum`), as well as the number of events

(dropout_count) standardized by the number of students enrolled in that course, that will perform as the mark of the counting process. Afterwards, these covariates are employed for the AG model for recurrent events describing the dropouts.

Compensators are then reconstructed as described in Eq. (2). In Figure 5, we display the baseline cumulative hazard and the reconstructed compensators, on two different scales. The behavior of the curves is notable: there is a steep increase in dropout counts at the beginning and end of the first year, particularly pronounced for specific degree courses. This pattern can be explained by several factors. Early in the first year, high dropout rates are often observed as students realize that the degree course they have chosen does not meet their expectations, leading them to switch programs or drop out. Additionally, many dropouts may occur by the end of the first year because students find the coursework too challenging or the degree program not aligned with their career aspirations. This combination of early and end-of-year dropouts contributes to the distinct peaks observed in the cumulative hazard curves. Notable is the case of a degree course in school sC.

Afterwards, the compensators are decomposed as in Eq. (4). The number of principal components for both levels is chosen by setting a proportion of variance explained equal to 0.99. As a result, two principal components are retained for both levels. At the higher hierarchical level (denoted as level 1, corresponding to the school), the eigenvalues obtained are $\hat{\lambda}_1^{(1)} = 74.747$ and $\hat{\lambda}_2^{(1)} = 0.434$, while at the lower hierarchical level (level 2, corresponding to the course), the eigenvalues are $\hat{\lambda}_1^{(2)} = 67.236$ and $\hat{\lambda}_2^{(2)} = 1.130$.

Figure 6 illustrates the first and second eigenfunctions for each of the two levels. To improve interpretability, as suggested in Ramsay and Silverman (2005), Figure 7 shows the mean compensator functions $\hat{\mu}(t)$ (solid black line) along with perturbation curves (red dashed lines for positive perturbations and blue dot-dashed lines for negative) representing the eigenfunctions within one standard deviation (i.e., the square roots of the eigenvalues) from the mean, based on the MFPCA performed on $\hat{\Lambda}_{ij}(t)$.

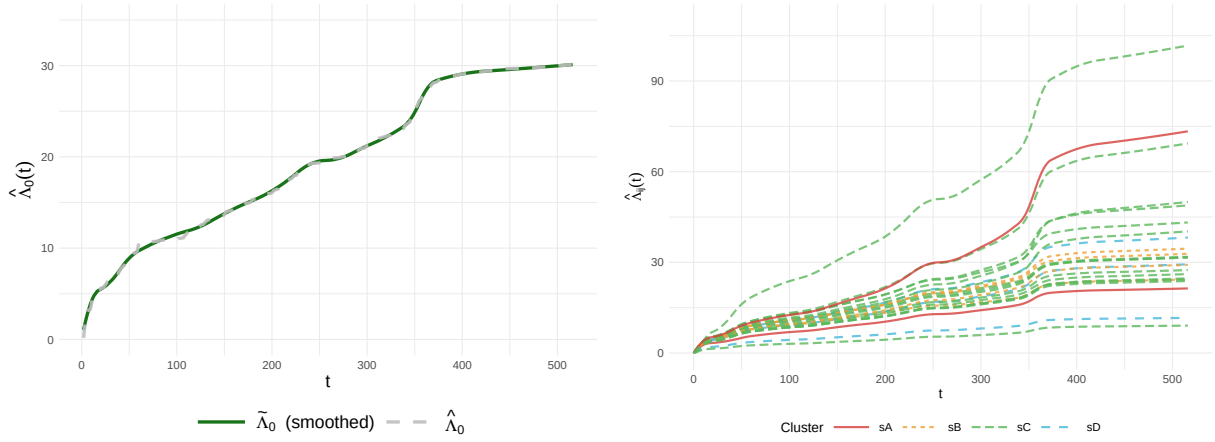
The distribution of dropouts over time reveals distinct patterns across schools and degree courses, each associated with varying dropout risks. Notably, our analysis, as a novel contribution to the existing literature, successfully disentangles the effects of schools from those of degree programs. The first principal components at both hierarchical levels capture deviations in dropout intensity relative to the average. Specifically, schools and degree courses with a high score on the first principal component (represented by the red dashed lines) are likely to experience a higher-than-average dropout rate, while those with a low score (blue dot-dashed lines) are likely to see fewer dropouts than average. Interestingly, the dropout patterns differ between the school and course levels: at the school level, there is a smoother increase in dropouts toward the end of the first year (second semester), likely due to fewer fluctuations compared to the course level, where more variation is observed.

The second principal components, though associated with less explained variance, highlight additional temporal contrasts. At the school level, institutions with a high score (red dashed curve) tend to experience fewer dropouts during the first two semesters but more dropouts in the third semester. This pattern is similarly observed at the course level, though with greater oscillations, suggesting that some noise may also be captured. Overall, these oscillations indicate more complex dropout dynamics at the course level, where factors influencing dropouts fluctuate more over time.

5.2 Predictive model

At this stage of the analysis, we are able to include the information derived from compensators into a predictive model (Subsection 5.2.1).

Before doing so, we first outline some data preprocessing steps necessary for preparing the dataset. We start by filtering the data to focus on career_start_ay = '2017'. To ensure consistency across observations, aligning with previous studies Masci et al. (2024); Ragni et al. (2024), we create the dichotomic variable age19 to denote whether students were above the age of 19 (1 if above 19, 0 otherwise). We compute the cumulative number of CFUs obtained by each student by the end of the first semester, in the variable ECTS1sem. The outcome variable dropout3y equals 1 if after first semester a student drops within three years, 0 otherwise. Following preprocessing, our dataset consists of 5666 students, of which 872 dropped out. Descriptive statistics for the over-described covariates after data pre-processing, are reported in Table 2, according to the dropout status. It is interesting to notice that, as expected, all numerical variables are higher when no dropout happens.



(i) Baseline cumulative hazard $\hat{\Lambda}_0(t)$ (ii) Reconstructed compensators $\hat{\Lambda}_{ij}(t)$

Figure 5: In panel (i) we show the baseline cumulative hazard of the AG model for recurrent events for the marked stochastic processes describing the dropouts. In panel (ii), we represent the reconstructed compensators as in (2) of the latter processes, each line representing a different degree `course` and each color and line type representing a different `school`, as indicated in legend.

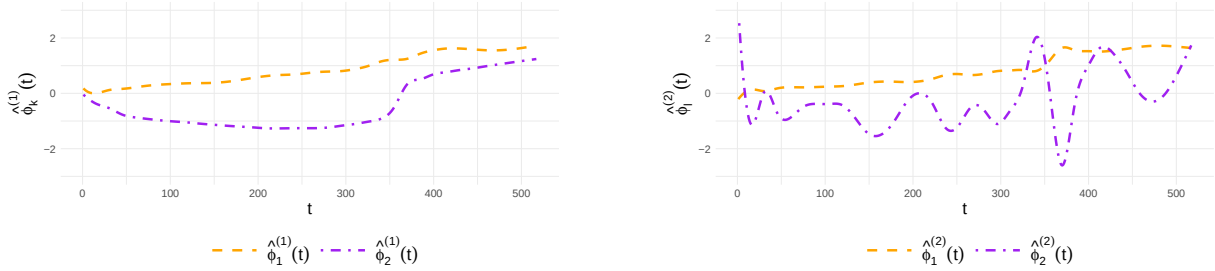
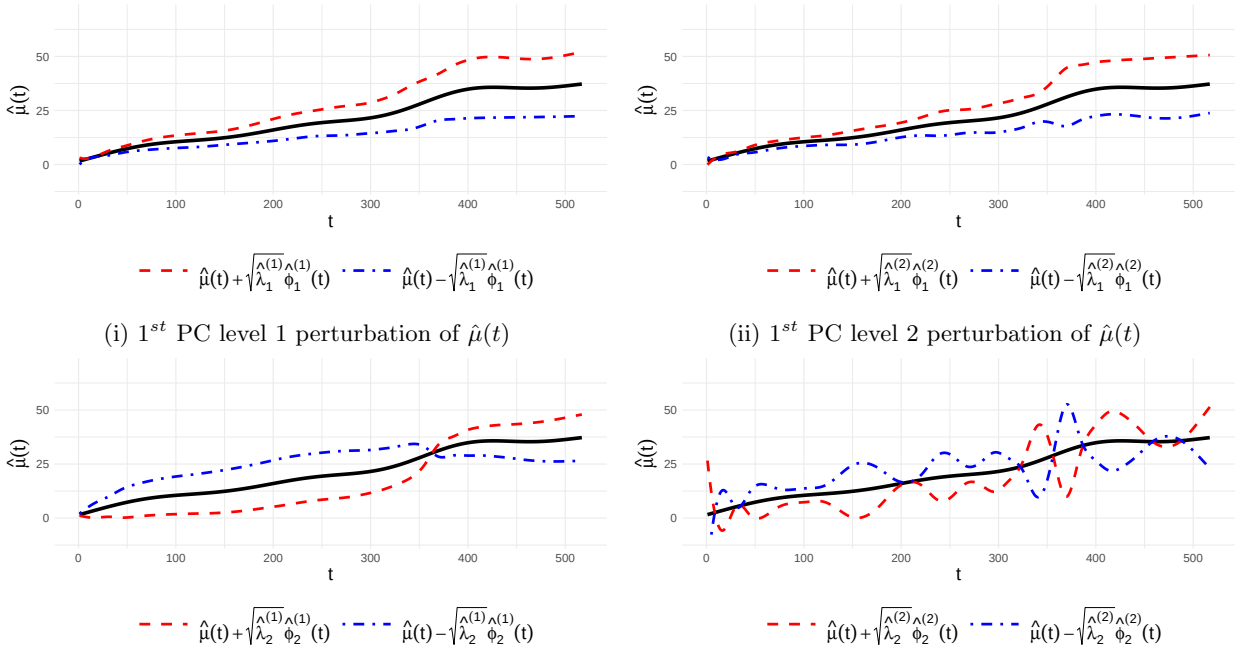


Figure 6: First two eigenfunctions at levels `school` and `course` levels (left and right panels, respectively), computed from the obtained $\hat{\Lambda}_{ij}(t)$.



(i) 1^{st} PC level 1 perturbation of $\hat{\mu}(t)$

(ii) 1^{st} PC level 2 perturbation of $\hat{\mu}(t)$

(iii) 2^{nd} PC level 1 perturbation of $\hat{\mu}(t)$

(iv) 2^{nd} PC level 2 perturbation of $\hat{\mu}(t)$

Figure 7: Average compensators curves $\hat{\mu}(t) = \frac{1}{n}\sum_{i,j}\hat{\Lambda}_{ij}(t)$ and their perturbations as indicated in legends, for 1^{st} principal component for `school` level in left panel (i) and `course` level in right panel (ii), and 2^{nd} principal component for `school` level in left panel (iii) and `course` level in right panel (iv).

Variable			`dropout3y=0`	`dropout3y=1`
Type	**Name**		**Mean (sd)**	**Mean (sd)**
Numerical	`admission_score`		67.00 (11.46)	63.61 (11.62)
	`ECTS1sem`		49.79 (14.74)	10.38 (15.39)
		Category	**N (Frequency)**	**N (Frequency)**
Categorical	`origins`	`OnSite`*	1033 (21.55%)	232 (26.61%)
		`Commuter`	3427 (71.49%)	583 (66.86%)
		`Offsite`	334 (6.96%)	57 (6.53%)
	`gender`	`Male`*	3200 (66.75%)	670 (76.83%)
		`Female`	31594 (33.25%)	202 (23.17%)
	`highschool_type`	`Scientific`*	3181 (66.36%)	542 (62.15%)
		`Classical`	327 (6.82%)	55 (6.31%)
		`Others`	607 (12.66%)	90 (10.32%)
		`Technical`	679 (14.16%)	185 (21.22%)
	`income`	`Medium`*	988 (20.61%)	125 (14.33%)
		`Grant`	1404 (29.29%)	287 (32.91%)
		`High`	1771 (36.94%)	378 (43.35%)
		`Low`	631 (13.16%)	82 (9.41%)
	`age19`	`0`*	4276 (89.19%)	671 (76.95%)
		`1`	518 (10.81%)	201 (23.05%)

* Reference category.

Table 2: Descriptive statistics for considered covariates after data pre-processing for `career_start_ay`='2017', according to the dropout status by the end of the third year.

5.2.1 Logistic regression model with functional compensators

We aim to model the probability of student dropout within 3 years after the first semester (`dropout3y`), using covariates at the `studentID`-level and functional principal component scores derived from the cumulative hazard of historical dropouts over time.

We consider two principal components ($K = 2$) at the `school` level, and one principal component ($L = 1$) at the `course` level, as there is low explained variability and high oscillations in the second component that could negatively affect model performance due to noise amplification. The binary outcome variable Y_{ijh} indicating whether a student h within course j and school i drops out within 3 years is modeled as $Y_{ijh} \sim \text{Bernoulli}(p_{ijh})$, linear predictor given by

$$\text{logit}(p_{ijh}) = \boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \sum_{k=1}^{K} \xi_{ik}\alpha_k^{(1)} + \sum_{l=1}^{L} \zeta_{ijl}\alpha_l^{(2)}$$

for $i = 1, ..., I$, $j = 1, ..., J_i$ and $h = 1, ...H_{ij}$. The vector of covariates $\mathbf{w}_{ijh}$ at the `studentID`-level includes demographic and academic information that could influence dropout risk, i.e., `origins`, `gender`, `highschool_type`, `income`, `age19`, `admission_score`, `ECTS1sem`. The choice of these variables is guided by previous literature, see for instance Masci et al. (2024). Obtained results for the estimated coefficients are reported in Table 3.

Parameter	Estimate	Std. Error	p-value
$\hat{\gamma}_0$ (Intercept)	0.589	0.029	0.000
$\hat{\gamma}_1$ (`origins` - Commuter)	0.022	0.008	0.008
$\hat{\gamma}_2$ (`origins` - Offsite)	-0.018	0.015	0.244
$\hat{\gamma}_3$ (`gender` – Female)	0.025	0.008	0.002
$\hat{\gamma}_4$ (`highschool_type` – Classical)	-0.000	0.014	0.988
$\hat{\gamma}_5$ (`highschool_type` – Others)	0.008	0.012	0.467
$\hat{\gamma}_6$ (`highschool_type` – Technical)	-0.002	0.010	0.812
$\hat{\gamma}_7$ (`income` – Grant)	-0.014	0.010	0.172
$\hat{\gamma}_8$ (`income` – High)	0.004	0.009	0.674
$\hat{\gamma}_9$ (`income` – Low)	-0.025	0.012	0.046
$\hat{\gamma}_{10}$ (`age19` – 1)	0.007	0.011	0.498
$\hat{\gamma}_{11}$ (`admission_score`)	0.001	0.000	0.000
$\hat{\gamma}_{12}$ (`ECTS1sem`)	-0.012	0.000	0.000
$\hat{\alpha}_1^{(1)}$ (ξ_{i1} - `school`-level score 1)	0.002	0.029	0.049
$\hat{\alpha}_2^{(1)}$ (ξ_{i2} - `school`-level score 2)	-0.070	0.001	0.000
$\hat{\alpha}_1^{(2)}$ (ζ_{ij1} - `course`-level score 1)	0.001	0.020	0.083

Table 3: Estimates, standard errors, and p-values for the logistic regression model with functional compensators.

It is interesting to notice that coefficients related to the obtained scores at `school`-level are significant and, specifically, the one related to the first principal component is positive, indicating that the probability of dropping out within the 3 years is increased if the school in which a student is enrolled has an high score on the first principal component, and this result is coherent with the plot in Figure 7 (i). On the other hand, the coefficient related to the the second principal component at the `school`-level is negative and statistically significant, indicating that students in schools with higher dropouts with respect to the average in the first year and lower than average in the third semester have higher probability to dropout. At the `course`-level, the first principal component is again positively associated with dropout risk, meaning students enrolled in courses where the dropout trend is above the average are linked to an increased probability of students dropping out.

Regarding the other covariates, `ECTS1sem` (credits earned in the first semester) is highly significant and negatively associated with dropout probability. This implies that students who pass more credits in their first semester have a lower risk of dropping out within three years. This result is in line with previous research, such as Masci et al. (2024) which highlights the strong predictive power of first-semester credits over later academic performance in determining dropout risk. The influence of first-semester credits is

particularly pronounced, as passing more credits early on seems to provide a greater protective effect against dropout.

For the remaining covariates, the results are largely consistent with expectations, although many are not statistically significant. For example, the coefficients related to `income` and `highschool_type` align with prior studies, but they lack statistical significance in this specific model. One exception is the variable `admission_score`, which is statistically significant but reveals an unexpected positive association with dropout probability. Typically, one would expect that a higher admission score is related with a reduced risk of dropout, as it often indicates greater preparedness for higher education. However, at PoliMi, students are permitted to take the admission test as early as their fourth year of secondary education. At this stage, they may not have fully developed the necessary competencies or maturity required for success in a university setting. Consequently, students who achieve high admission scores at this early stage may still struggle academically once enrolled, potentially increasing their likelihood of dropping out. This surprising result is the focus of ongoing study at PoliMi, as we aim to better understand the underlying factors contributing to this unexpected association.

In terms of model performance, the model achieves an AIC (Akaike, 1998; Bozdogan, 1987) of 808.68, and excellent predictive power, with an AUC (Hanley and McNeil, 1982) of 0.9425 and an accuracy of 0.92. Sensitivity (0.954) and precision (0.951) are both high, indicating that the model is very good at correctly identifying students who are at risk of dropping out. Specificity (0.732), though slightly lower, still indicates a reasonable ability to identify students who are not at risk. These performance metrics suggest that the model is well-calibrated for predicting dropout risk, with a particular strength in identifying students at higher risk.

If we fit the same model but exclude the compensators' information, we obtain an AIC of 823.85 and an AUC of 0.9418. This indicates that the model provides a better fit when compensator information is included. On the other hand, while a mixed-effects model can account for unobserved heterogeneity, it primarily introduces scaling factors and may not be as suitable in our case. It lacks the capacity to capture the intricate temporal dynamics that our compensator-based approach, combined with multilevel functional principal component analysis, effectively models over time. Also, the authors in Baraldo et al. (2013) compared such models, demonstrating that mixed-effects models do not offer superior performance.

6 Discussion

Addressing student dropouts is a critical concern for universities, both academically and financially. Each dropout represents an inefficient use of institutional resources allocated to recruitment, teaching, and student support. Reducing dropout rates directly impacts both financial stability and the overall effectiveness of educational systems.

One of the complexities in tackling this issue lies in the heterogeneous nature of dropout behaviour across degree programs and schools. Different academic disciplines present unique challenges - some programs may experience high dropout rates early on due to demanding foundational courses, while others see increased dropouts as students' careers progress. Similarly, the dropout patterns can vary considerably across schools within the same university, influenced by factors such as faculty engagement and available student support.

In this paper, we present a novel approach to modelling dropout behaviour by examining occurrences over time within both degree programs and schools. Our work has two main goals: (i) to estimate the dropout trends over time and examine its variability across different degree programs and schools, and (ii) to leverage this information in a predictive framework at the student level. To achieve these objectives, we utilize Cox-based regression for recurrent events to capture the temporal dynamics and underlying structure of dropout trends. In this initial phase, we employ an AG model, as supported by existing literature (Spreafico and Ieva, 2021). However, it is important to note that other modeling choices, such as those proposed by Baraldo et al. (2013), which build on Peña et al. (2007), are also possible. Selecting the appropriate model can be challenging; consequently, this first step of the analysis could be replicated using alternative modeling approaches, allowing for further exploration and validation of our findings. By decomposing dropout patterns within programs and schools through multilevel functional principal component analysis, we provide a detailed view of critical time periods when dropout rates tend to spike. This approach offers both visual and quantitative insights into when students are most at risk of leaving their studies, allowing institutions to identify vulnerable cohorts and periods. Furthermore, by capturing these temporal trends, we gain a deeper understanding of how dropout behaviour varies across disciplines.

Our predictive model adds significant value by incorporating historical dropout data on current dropout behaviour. By integrating information from previous cohorts, our approach allows universities to more accurately forecast future dropout risks and target proactive interventions. This enables educational institutions to identify at-risk students earlier in their academic journeys, based on a combination of baseline characteristics such as academic performance in first semester, socioeconomic status, and previous schools attended. With this information, universities can implement more personalized support strategies aimed at reducing dropouts, such as tutoring classes, thus improving student retention and overall success.

While our model presents promising results, there are several limitations and areas for further development. First, this is a preliminary analysis, and the results should be validated across multiple academic years to ensure robustness. Cross-validation techniques could be employed to improve the stability and generalizability of the model outcomes. Additionally, the impact of external factors like the Covid-19 pandemic - which may have fundamentally altered student engagement and retention - should be incorporated into future analyses. Understanding how the pandemic influenced dropout patterns could further refine our predictions. Furthermore, our analysis focuses on the first three semesters, a period selected because highly predictive of dropout risk. Since the compensator must be integrated over historical data, extending the observation period beyond this point did not yield significant improvements in the analysis; however, this remains an area for further evaluation in future research. Indeed, while considering this period in the compensators reconstruction allows capturing early dropouts at course and school levels, future extensions could consider the entire three-year duration of undergraduate programs to provide a more comprehensive understanding of student retention. Incorporating time-to-event data would allow us to model dropout risk more accurately over time, addressing both whether and when students are likely to drop out.

Competing interests

No competing interest is declared.

Data availability

The participants of this study did not give written consent for their data to be shared publicly, so due to the sensitive nature of the research, the full supporting data is not available.

Acknowledgements

The authors acknowledge the support by MUR, Italy, grant 'Dipartimento di Eccellenza 2023-2027'.

A Compensators reconstruction

The realizations of each compensator $\Lambda_{ij}(t)$ for each unit j in cluster i, employing result in Eq. (1) can be expressed as follows:

$$\begin{aligned}
\Lambda_{ij}(t) &= \int_0^t Y_{ij}(s)\lambda_0(s)\,\exp(\boldsymbol{\beta}^T\mathbf{x}_{ij}(s)+\boldsymbol{\theta}^T\mathbf{z}_{ij}(s))ds \\
&= \sum_{k=0}^{N_{ij}(t^-)} \int_{t_k^{(ij)}}^{t_{k+1}^{(ij)}\wedge t} \lambda_0(s)\;\exp(\boldsymbol{\beta}^T\mathbf{x}_{ij}(s)+\boldsymbol{\theta}^T\mathbf{z}_{ij}(s))ds \\
&\simeq \sum_{k=0}^{N_{ij}(t^-)} \exp(\boldsymbol{\beta}^T\mathbf{x}_{ij}(t_k^{(ij)})+\boldsymbol{\theta}^T\mathbf{z}_{ij}(t_k^{(ij)})) \int_{t_k^{(ij)}}^{t_{k+1}^{(ij)}\wedge t} \lambda_0(s)ds \\
&= \sum_{k=0}^{N_{ij}(t^-)} \exp(\boldsymbol{\beta}^T\mathbf{x}_{ij}(t_k^{(ij)})+\boldsymbol{\theta}^T\mathbf{z}_{ij}(t_k^{(ij)}))\big[\tilde{\Lambda}_0(t_{k+1}^{(ij)}\wedge t)-\tilde{\Lambda}_0(t_k^{(ij)})\big]
\end{aligned}$$

where $a\wedge b=\min\{a,b\}$, $N_{ij}(t^-)$ represents the number of occurrences that have happened strictly before time t, and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are the estimated vectors of coefficients.

References

Aalen, O., O. Borgan, and H. Gjessing (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aina, C., E. Baici, G. Casalone, and F. Pastore (2018). The economics of university dropouts and delayed graduation: a survey.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

Amorim, L. D. and J. Cai (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology 44*(1), 324–333.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

Andersen, P. K. and N. Keiding (2002). Multi-state models for event history analysis. *Statistical methods in medical research 11*(2), 91–115.

Arulampalam, W., R. A. Naylor, and J. P. Smith (2004). A hazard model of the probability of medical school drop-out in the uk. *Journal of the Royal Statistical Society Series A: Statistics in Society 167*(1), 157–178.

Atzeni, G., L. G. Deidda, M. Delogu, and D. Paolini (2022). Drop-out decisions in a cohort of italian universities. In *Teaching, Research and Academic Careers: An Analysis of the Interrelations and Impacts*, pp. 71–103. Springer International Publishing Cham.

Baraldo, S., F. Ieva, A. M. Paganoni, and V. Vitelli (2013). Outcome prediction for heart failure telemonitoring via generalized linear models with functional covariates. *Scandinavian Journal of Statistics 40*(3), 403–416.

Beutner, E. (2023). A review of effective age models and associated non-and semiparametric methods. *Econometrics and Statistics 28*, 105–119.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika 52*(3), 345–370.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 45–57.

Cannistrà, M., C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni (2022). Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education 47*(9), 1935–1956.

Cannistrà, M. (2024). *Reducing dropout rates: the challenge of learning analytics in higher education institutions*. Ph. D. thesis, Politecnico di Milano.

Cook, R. J., J. F. Lawless, et al. (2007). *The statistical analysis of recurrent events*. Springer.

Cui, E., R. Li, C. M. Crainiceanu, and L. Xiao (2023). Fast multilevel functional principal component analysis. *Journal of Computational and Graphical Statistics 32*(2), 366–377.

Daley, D. J. and D. Vere-Jones (2002). *An Introduction to the Theory of Point Processes*. Springer New York, NY.

Daley, D. J., D. Vere-Jones, et al. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Di, C.-Z., C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi (2009). Multilevel functional principal component analysis. *The annals of applied statistics 3*(1), 458.

Diao, L., R. J. Cook, and K.-A. Lee (2014). Statistical analysis of recurrent adverse events. *Statistical Methods for Evaluating Safety in Medical Product Development*, 180–192.

Diaz Lema, M., M. Vooren, M. Cannistrà, C. van Klaveren, T. Agasisti, and I. Cornelisz (2024). Predicting dropout in higher education across borders. *Studies in Higher Education 49*(1), 141–156.

Gury, N. (2011). Dropping out of higher education in france: a micro-economic approach using survival analysis. *Education Economics 19*(1), 51–64.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*(1), 29–36.

Karhunen, K. (1947). Under lineare methoden in der wahr scheinlichkeitsrechnung. *Annales Academiae Scientiarun Fennicae Series A1: Mathematia Physica 47*.

Katz, L. F. and K. M. Murphy (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics 107*(1), 35–78.

Kijima, M., H. Morimura, and Y. Suzuki (1988). Periodical replacement problem without assuming minimal repair. *European Journal of Operational Research 37*(2), 194–203.

Kleinbaum, D. G. and M. Klein (1996). *Survival analysis a self-learning text*. Springer.

Last, G. and A. Brandt (1995). *Marked Point Processes on the real line: the dynamical approach*. Springer Science & Business Media.

Lin, D. Y., L.-J. Wei, I. Yang, and Z. Ying (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 711–730.

Loeve, M. (1948). Functions aleatoires du second ordre. *Processus stochastique et mouvement Brownien*, 366–420.

Masci, C., M. Cannistrà, and P. Mussida (2024). Modelling time-to-dropout via shared frailty cox models. a trade-off between accurate and early predictions. *Studies in Higher Education 49*(4), 763–781.

Meyer, P.-A. (1962). A decomposition theorem for supermartingales. *Illinois Journal of Mathematics 6*(2), 193–205.

Min, Y., G. Zhang, R. A. Long, T. J. Anderson, and M. W. Ohland (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of Engineering Education 100*(2), 349–373.

Morris, J. S., M. Vannucci, P. J. Brown, and R. J. Carroll (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association 98*(463), 573–583.

Mussida, P. and P. L. Lanzi (2022). A computational tool for engineer dropout prediction. In *2022 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1571–1576. IEEE.

OECD (2019). Education at a glance 2019.

Pasupathy, R. (2010). Generating homogeneous poisson processes. *Wiley encyclopedia of operations research and management science*.

Patacsil, F. F. (2020). Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. *Universal Journal of Educational Research 8*(9), 4036–4047.

Peña, E. A. and M. Hollander (2004). Models for recurrent events in reliability and survival analysis. *Mathematical reliability: An expository perspective*, 105–123.

Peña, E. A., E. H. Slate, and J. R. González (2007). Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference 137*(6), 1727–1747.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Plank, S. B., S. DeLuca, and A. Estacion (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education 81*(4), 345–370.

Prentice, R. L., B. J. Williams, and A. V. Peterson (1981). On the regression analysis of multivariate failure time data. *Biometrika 68*(2), 373–379.

Ragni, A., D. Ippolito, and C. Masci (2024). Assessing the impact of hybrid teaching on students' academic performance via multilevel propensity score-based techniques. *Socio-Economic Planning Sciences 92*, 101824.

Ramsay, J. and B. Silverman (2005). Principal components analysis for functional data. *Functional data analysis*, 147–172.

Romani, G. (2023). Time-varying shared frailty cox models for the analysis of university students dropout. Master's thesis, Politecnico di Milano.

Spreafico, M. and F. Ieva (2021). Functional modeling of recurrent events on time-to-event processes. *Biometrical Journal 63*(5), 948–967.

Therneau, T. M., P. M. Grambsch, T. M. Therneau, and P. M. Grambsch (2000). *The Cox model*. Springer.

Tinto, V. (1982). Defining dropout: A matter of perspective. *New Directions for Institutional Research 1982*(36), 3–15.

Vallejos, C. A. and M. F. Steel (2017). Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society Series A: Statistics in Society 180*(2), 613–631.

Wei, L.-J., D. Y. Lin, and L. Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association 84*(408), 1065–1073.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

98/2024 Castiglione, C.; Arnone, E.; Bernardi, M.; Farcomeni, A.; Sangalli, L.M.
PDE-regularised spatial quantile regression

97/2024 Ferro, N.; Mezzadri, F.; Carbonaro, D.; Galligani, E.; Gallo, D.; Morbiducci, U.; Chiastra, C.; Perotto, S.
Designing novel vascular stents with enhanced mechanical behavior through topology optimization of existing devices

96/2024 Brivio, S.; Fresca, S.; Manzoni, A.
PTPI-DL-ROMs: Pre-trained physics-informed deep learning-based reduced order models for nonlinear parametrized PDEs

93/2024 Conti, P.; Kneifl, J.; Manzoni, A.; Frangi, A.; Fehr, J.; Brunton, S.L.; Kutz, J.N.
VENI, VINDy, VICI - a variational reduced-order modeling framework with uncertainty quantification

94/2024 Franco, N.R.; Fresca, S.; Tombari, F.; Manzoni, A.
Deep Learning-based surrogate models for parametrized PDEs: handling geometric variability through graph neural networks

95/2024 Zacchei, F.; Rizzini, F.; Gattere, G.; Frangi, A.; Manzoni, A.
Neural networks based surrogate modeling for efficient uncertainty quantification and calibration of MEMS accelerometers

91/2024 Ciaramella, G.; Kartmann, M.; Mueller, G.
Solving Semi-Linear Elliptic Optimal Control Problems with L1-Cost via Regularization and RAS-Preconditioned Newton Methods

Castiglionea, C.; Arnonec, E.; Bernardi, M.; Farcomeni, A.; Sangalli, L.M.
PDE regularised spatial quantile regression

88/2024 Regazzoni, F.; Poggesi, C.; Ferrantini, C.
Elucidating the cellular determinants of the end-systolic pressure-volume relationship of the heart via computational modelling

85/2024 Brivio, S.; Franco, Nicola R.; Fresca, S.; Manzoni, A.
Error estimates for POD-DL-ROMs: a deep learning framework for reduced order modeling of nonlinear parametrized PDEs enhanced by proper orthogonal decomposition